

Semiparametric inference in mixture models with predictive recursion marginal likelihood

Ryan Martin

Department of Mathematical Sciences
Indiana University–Purdue University Indianapolis
rgmartin@math.iupui.edu

Surya T. Tokdar

Department of Statistical Science
Duke University
tokdar@stat.duke.edu

April 7, 2011

Abstract

Predictive recursion is an accurate and computationally efficient algorithm for nonparametric estimation of mixing densities in mixture models. In semiparametric mixture models, however, the algorithm fails to account for any uncertainty in the additional unknown structural parameter. As an alternative to existing profile likelihood methods, we treat predictive recursion as a filter approximation to fitting a fully Bayes model, whereby an approximate marginal likelihood of the structural parameter emerges and can be used for inference. We call this the predictive recursion marginal likelihood. Convergence properties of predictive recursion under model mis-specification also lead to an attractive construction of this new procedure. We show pointwise convergence of a normalized version of this marginal likelihood function. Simulations compare the performance of this new marginal likelihood approach that of existing profile likelihood methods as well as Dirichlet process mixtures in density estimation. Mixed-effects models and an empirical Bayes multiple testing application in time series analysis are also considered.

Keywords and phrases. Density estimation; Dirichlet process mixture; empirical Bayes; filtering algorithm; marginal likelihood; martingale; mixed effects model; multiple testing; profile likelihood.

1 Introduction

Consider data Y_1, \dots, Y_n modeled as independent draws from a common, nonparametric mixture distribution with density

$$m_f(y) = \int p(y | u) f(u) d\mu(u), \quad y \in \mathcal{Y}, \quad (1)$$

where $(y, u) \mapsto p(y | u)$ is a known kernel on $\mathcal{Y} \times \mathcal{U}$ and f is an unknown mixing density in \mathbb{F} , the set of densities with respect to a σ -finite Borel measure μ on \mathcal{U} . Newton et al. (1998) introduced the following stochastic algorithm, called predictive recursion, to estimate f and m_f .

PR Algorithm. Choose an initial estimate $f_0 \in \mathbb{F}$ of f , and a sequence of weights $w_1, \dots, w_n \in (0, 1)$. For $i = 1, \dots, n$, compute the following:

$$f_i(u) = (1 - w_i)f_{i-1}(u) + w_i \frac{p(Y_i | u)f_{i-1}(u)}{\int p(Y_i | u')f_{i-1}(u') d\mu(u')}, \quad u \in \mathcal{U}, \quad (2)$$

$$m_i(y) = m_{f_i}(y) = \int p(y | u)f_i(u) d\mu(u), \quad y \in \mathcal{Y}. \quad (3)$$

Return f_n and m_n as the estimates of f and m_f , respectively.

The algorithm's strengths include its fast computation and its unique flexibility to estimate the mixing density with respect to any user-specified dominating measure μ . Predictive recursion also has a close connection to Dirichlet process mixture models; see Newton et al. (1998), Quintana and Newton (2000), Newton and Zhang (1999), Newton (2002), and Section 2. Tokdar et al. (2009) show that when Y_1, \dots, Y_n are generated independently from a density m which equals m_{f^*} for some $f^* \in \mathbb{F}$, the resulting estimates f_n and m_n converge as $n \rightarrow \infty$, respectively and in appropriate topologies, to f^* and m_{f^*} ; see also Ghosh and Tokdar (2006) and Martin and Ghosh (2008). Martin and Tokdar (2009) show that if m does not equal m_f for any $f \in \mathbb{F}$, then the estimates still converge, but now the limits are characterized by the minimizer f^* of the Kullback–Leibler divergence $K(m, m_f) = \int \log\{m(y)/m_f(y)\}m(y) dy$. The minimizer exists and is unique under certain conditions. An upper bound on the rate of this convergence is also available.

In statistical applications, however, an exact description of the kernel $p(y | u)$ is rarely available. It is more common to use a family of kernels $p(y | \theta, u)$ indexed by a parameter $\theta \in \Theta$ and model Y_1, \dots, Y_n as independent draws from a semiparametric mixture

$$m_{f,\theta}(y) = \int p(y | \theta, u)f(u) d\mu(u), \quad (4)$$

where both θ and f are unknown. A frequently encountered example of this is density estimation with mixtures of Gaussian kernels, where $p(y | \theta, u) = N(y | u, \theta^2)$ with θ playing the role of a bandwidth. A related formulation is in the linear mixed effects model $Y_i = U_i + X_i'\beta + \sigma\varepsilon_i$, where $\theta = (\beta, \sigma)$ is unknown, and the density f of the random effect U_i is not restricted to a parametric family. While θ is more like a nuisance parameter in the density estimation problem, it takes center stage in the mixed effect model. In either case, predictive recursion fails to provide any statistical analysis for θ .

Tao et al. (1999) counter this shortcoming by embedding predictive recursion in a profile likelihood framework. At any given θ , one runs the predictive recursion algorithm with kernel $p(y | \theta, u)$ and a suitable initial guess $f_{0,\theta}$ to recursively compute $f_{i,\theta}$ and $m_{i,\theta}$ for $i = 1, \dots, n$. The final update $m_{n,\theta}$ is then plugged in to give the following profile likelihood in θ :

$$L_n^p(\theta) = \prod_{i=1}^n m_{n,\theta}(Y_i). \quad (5)$$

Tao et al. (1999) maximize this profile likelihood to estimate θ . Such a plug-in approach does not account for the lack of precision in estimating the mixing density f . In the density estimation setting, the profile likelihood may be maximized at the zero bandwidth, completely ignoring the extreme variability of the estimates of f at small bandwidths. Such undesirable behavior can be avoided by imposing a penalty on the estimate of f . But a general framework along these lines is yet to emerge, particularly for problems where inference on θ is the main focus.

In this paper we demonstrate that predictive recursion's close connection with the Bayesian paradigm offers a rich alternative to the plug-in approach. By viewing it as an approximation to fitting a fully Bayesian model on (θ, f) , it is natural to ask whether it can also provide an approximation to the marginal likelihood for θ as defined by the Bayesian model. In Section 2 we show that such an approximation is indeed available and of the form

$$L_n^M(\theta) = \prod_{i=1}^n m_{i-1,\theta}(Y_i). \quad (6)$$

The approximate marginal likelihood $L_n(\theta)$, which we call the predictive recursion marginal likelihood, appears to inherit the intrinsic Ockham's razor properties (Jefferys and Berger 1992) of the original Bayesian formulation. That is, the θ values for which the conditional prior on f is more spread out automatically receive greater penalty.

In Section 3 we show that if Y_1, \dots, Y_n are independent samples from a density m , then $\log L_n^M(\theta)$ equals $-n \inf_{f \in \mathbb{F}} K(m, m_{f,\theta})$ plus a quantity that grows slower than n . A consequence of this is a convergence property of the maximum predictive recursion marginal likelihood estimate

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} L_n^M(\theta) \quad (7)$$

that follows from an argument similar to that of Wald (1949). Specifically, if Θ is finite, then $\hat{\theta}_n$ converges to θ^* as $n \rightarrow \infty$, where the limit is characterized by the minimizer (θ^*, f^*) of $K(m, m_{f,\theta})$ over $\Theta \times \mathbb{F}$. Our simulation studies suggest that similar results should hold for compact Θ as well, but so far a proof has eluded us.

An exact sampling distribution for $\hat{\theta}_n$ in (7) is not available. Therefore, for inference on θ we estimate the standard error of $\hat{\theta}_n$ via the curvature of L_n^M at its maxima. This is motivated by the interpretation of the predictive recursion marginal likelihood as an approximate Bayesian marginal likelihood for which Laplace approximation applies (Tierney and Kadane 1986).

Several examples are presented in Section 4. For density estimation, our simulations indicate that L_n^M closely approximates Bayes Dirichlet process mixture marginal likelihood, whereas L_n^P is more sporadic, in some cases concentrating on the boundary of the parameter space. Applications to interval estimation in random-intercept regression models and multiple testing in mixtures of autoregressive process models are also given.

2 Approximation to the Dirichlet process mixture marginal likelihood

As noted in Newton et al. (1998) and Newton (2002), the updating scheme (2) has a close connection with the posterior updates in a Bayesian formulation when f is modeled by

a Dirichlet process prior. In this section we further explore this connection to establish $L_n(\theta)$ as an approximation to the marginal likelihood of θ as defined by such a Bayesian formulation.

To be precise, consider the following extension of the mixture model (4):

$$m(y) = m_{F,\theta}(y) = \int p(y \mid \theta, u) dF(u), \quad (8)$$

where F is an unknown probability measure on \mathcal{U} , not necessarily dominated by μ . Consider a Bayesian formulation

$$Y_1, \dots, Y_n \mid (F, \theta) \stackrel{\text{iid}}{\sim} m_{F,\theta}, \quad F \mid \theta \sim \Pi_\theta, \quad \theta \sim \Gamma, \quad (9)$$

where Π_θ is, for each $\theta \in \Theta$, a probability distribution over the space of probability measures F , and Γ is a probability distribution on Θ . The posterior distribution Γ_n of θ given the n observations can be written as $d\Gamma_n(\theta) \propto L_n^B(\theta) d\Gamma(\theta)$, where

$$L_n^B(\theta) = \int \left\{ \prod_{i=1}^n m_{F,\theta}(Y_i) \right\} d\Pi_\theta(F)$$

is the marginal likelihood of θ obtained by integrating out F from (9). For every $\theta \in \Theta$, let $\Pi_{i,\theta}$ denote the conditional posterior distribution of F given θ and the first i observations, i.e., $d\Pi_{i,\theta}(F) \propto \left\{ \prod_{j=1}^i m_{F,\theta}(Y_j) \right\} d\Pi_\theta(F)$. Then by linearity of $m_{F,\theta}$ and Fubini's theorem,

$$\begin{aligned} L_n^B(\theta) &= \prod_{i=1}^n \int m_{F,\theta}(Y_i) d\Pi_{i-1,\theta}(F) = \prod_{i=1}^n \int p(Y_i \mid \theta, u) dF(u) d\Pi_{i-1,\theta}(F) \\ &= \prod_{i=1}^n \int p(Y_i \mid \theta, u) dF_{i-1,\theta}(u), \end{aligned} \quad (10)$$

where $F_{i,\theta} = \int F d\Pi_{i,\theta}(F)$ is the conditional posterior mean of F given $(Y_1, \dots, Y_i, \theta)$.

Now consider the special case where $\Pi_\theta = \text{DP}(\alpha_0, F_{0,\theta})$, the Dirichlet process distribution with precision parameter $\alpha_0 > 0$ and base measure $F_{0,\theta}$ (Ferguson 1973; Ghosh and Ramamoorthi 2003). Assume that the base measures $F_{0,\theta}$ are all absolutely continuous with respect to μ , admitting densities $f_{0,\theta} = dF_{0,\theta}/d\mu \in \mathbb{F}$. It follows from the Polya urn representation of a Dirichlet process (Blackwell and MacQueen 1973) that

$$dF_{1,\theta}(u) = \frac{\alpha}{\alpha + 1} dF_{0,\theta}(u) + \frac{1}{\alpha + 1} \frac{p(Y_1 \mid \theta, u) dF_{0,\theta}(u)}{\int p(Y_1 \mid \theta, u') dF_{0,\theta}(u')}. \quad (11)$$

Therefore, $F_{1,\theta}$ is absolutely continuous with respect to μ and the density $dF_{1,\theta}/d\mu \in \mathbb{F}$ is identical to the predictive recursion output $f_{1,\theta}$ based on the single observation Y_1 , with initial guess $f_{0,\theta}$, kernel $p(y \mid \theta, u)$ and weight $w_1 = 1/(1 + \alpha_0)$. Consequently $L_2^B(\theta) = L_2^M(\theta)$ as can be verified by comparing (6) and (10).

This analogy, however, does not carry over to $L_i^B(\theta)$ and $L_i^M(\theta)$ for $i \geq 3$. For $i = 3$, the relevant conditional posterior mean $F_{2,\theta}$ does not admit a representation as in (11) in terms of $F_{1,\theta}$ and $p(Y_2 \mid \theta, u)$ because the conditional posterior distribution $\Pi_{1,\theta}$ is no longer a Dirichlet process distribution, but rather a mixture of Dirichlet processes (Antoniak 1974).

To remedy this, consider an approximation to the Bayesian model, where we successively replace $\Pi_{i,\theta}$ with $\hat{\Pi}_{i,\theta} = \text{DP}(\alpha_i, \hat{F}_{i,\theta})$ where $\hat{F}_{0,\theta} = F_{0,\theta}$ and

$$\hat{F}_{i,\theta} = \frac{\int F m_{F,\theta}(Y_i) d\hat{\Pi}_{i-1,\theta}(F)}{\int m_{F,\theta}(Y_i) d\hat{\Pi}_{i-1,\theta}(F)}, \quad i \geq 1,$$

is what one would obtain for $E(F \mid Y_1, \dots, Y_i, \theta)$ if the conditional posterior of F given $(Y_1, \dots, Y_{i-1}, \theta)$ was indeed $\hat{\Pi}_{i-1,\theta}$. These successive replacements can be thought of as a dynamic, mean preserving, filter approximation to the original Bayesian model. Note that every $\hat{F}_{i,\theta}$ remains absolutely continuous with respect to μ and satisfies the recursion

$$d\hat{F}_{i,\theta}(u) = \frac{\alpha_{i-1}}{1 + \alpha_{i-1}} d\hat{F}_{i-1,\theta}(u) + \frac{1}{1 + \alpha_{i-1}} \frac{p(Y_i \mid \theta, u) d\hat{F}_{i-1,\theta}(u)}{\int p(Y_i \mid \theta, u') d\hat{F}_{i-1,\theta}(u')}.$$

This, coupled with the initial condition $\hat{F}_{0,\theta} = F_{0,\theta}$, implies that the densities $d\hat{F}_{i,\theta}/d\mu$ are precisely the $f_{i,\theta}$ that result from predictive recursion applied to the observations Y_1, \dots, Y_n , with initial guess $f_{0,\theta}$, kernel $p(y \mid \theta, u)$ and weights $w_i = 1/(1 + \alpha_{i-1})$. Therefore the corresponding approximation $\hat{L}_n^B(\theta) = \prod_{i=1}^n \int p(Y_i \mid \theta, u) d\hat{F}_{i-1,\theta}(u)$ of $L_n^B(\theta)$ is exactly $L_n^M(\theta)$.

For every $\theta \in \Theta$, the quantity $\prod_{i=1}^n m_{i-1,\theta}(Y_i)$ indeed defines a joint probability density for (Y_1, \dots, Y_n) which admits $L_n^M(\theta)$ as an exact likelihood function for θ . However it is unknown whether this joint density corresponds to any exchangeable hierarchical model on the Y_i 's, thus making it somewhat unsuitable to use it for statistical analysis. For this reason, we do not focus on studying $L_n^M(\theta)$ from the perspective of the joint model for which it is an exact likelihood function. We are instead interested in studying it as an inferential tool when Y_i 's are generated independently from a common density m which may or may not be a mixture as in (4).

3 Asymptotic theory

3.1 Notation and preliminaries

For \mathbb{F} , the set of all densities on \mathcal{U} with respect to μ as in Section 1, let $\overline{\mathbb{F}}$ be its closure with respect to the weak topology. With a slight abuse of notation, the elements of $\overline{\mathbb{F}}$ are also denoted by f , although they need not admit a density with respect to μ . For each θ , let $\mathbb{M}_\theta = \{m_{f,\theta} : f \in \overline{\mathbb{F}}\}$. Martin and Tokdar (2009) show that the predictive recursion estimates $m_{n,\theta}$ converge, for each fixed θ , to the best mixture density in \mathbb{M}_θ , if the following assumptions hold.

Assumption 1. Observations Y_1, Y_2, \dots are independent with a common density m , and $K(m, m')$ is finite for all $m' \in \mathbb{M} = \bigcup_{\theta \in \Theta} \mathbb{M}_\theta$.

Assumption 2. The weight sequence satisfies $\sum_n w_n = \infty$ and $\sum_n w_n^2 < \infty$.

Assumption 3. The set $\overline{\mathbb{F}}$ is compact with respect to the weak topology.

Assumption 4. The mapping $u \mapsto p(y \mid \theta, u)$ is bounded and continuous for all (y, θ) .

Assumption 5. For each (θ_1, θ_2) pair, there exists $A = A(\theta_1, \theta_2) < \infty$ such that

$$\sup_{u_1, u_2} \int \left\{ \frac{p(y|\theta_1, u_1)}{p(y|\theta_2, u_2)} \right\}^2 m(y) dy \leq A.$$

Assumption 2 is standard in the literature on stochastic approximation algorithms, of which predictive recursion is a special case (Martin and Ghosh 2008), and it holds if w_n decays like $n^{-\gamma}$ for $\gamma \in (1/2, 1]$. Assumption 3 is satisfied if, for example, \mathcal{U} is compact and μ is Lebesgue measure. The more demanding Assumption 5, holds for many standard kernels $p(y | \theta, u)$, such as those arising from Gaussian or other exponential family distributions, whenever \mathcal{U} is compact and m admits a moment-generating function on \mathcal{U} .

Define the mapping

$$K^*(\theta) = \inf \{ K(m, m_{f,\theta}) : f \in \overline{\mathbb{F}} \}, \quad \theta \in \Theta, \quad (12)$$

the smallest Kullback–Leibler divergence over \mathbb{M}_θ . Attainment of the infimum in (12) follows from Assumptions 3 and 4; see Martin and Tokdar (2009), Lemma 3.1. Let $a_n = \sum_{i=1}^n w_i$ denote the partial sums of the weight sequence $\{w_n : n \geq 1\}$. For two real sequences $\{\alpha_n\}$ and $\{\beta_n\}$, we write $\alpha_n = O(\beta_n)$ if α_n/β_n is bounded, and $\alpha_n = o(\beta_n)$ if $\alpha_n/\beta_n \rightarrow 0$. Then Martin and Tokdar (2009) prove a version of the following theorem.

Theorem 1. *Under Assumptions 1–5, $K(m, m_{n,\theta}) \rightarrow K^*(\theta)$ almost surely as $n \rightarrow \infty$ for each fixed θ . In addition to Assumptions 1–5, if the infimum in (12) is attained in the interior of \mathbb{F} , and if $\sum_n a_n w_n^2 < \infty$, then $K(m, m_{n,\theta}) - K^*(\theta) = o(a_n^{-1})$ almost surely.*

For weights that satisfy $w_n = O(n^{-\gamma})$, the extra condition, $\sum_n a_n w_n^2 < \infty$, in the second part of the theorem holds if and only if $\gamma \in (2/3, 1]$. Therefore, the best available rate in this case is $o(n^{-1/3})$ almost surely. But Martin and Tokdar (2009) argue that this rate is conservative.

3.2 Main results

Write $\ell_n = \log L_n^M$ and define the following normalization:

$$K_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log \frac{m(Y_i)}{m_{i-1,\theta}(Y_i)} = -\frac{\ell_n(\theta) - \ell_{0n}}{n}, \quad (13)$$

where $\ell_{0n} = \sum_{i=1}^n \log m(Y_i)$ is the log joint density of Y_1, \dots, Y_n . We will show that the main result of Theorem 1 holds with $K_n(\theta)$ in place of $K(m, m_{n,\theta})$.

Assumption 6. For each θ , there exists $B = B_\theta < \infty$, such that the density $m_{f,\theta} \in \mathbb{M}_\theta$ at which the infimum in (12) is attained satisfies

$$\int \left\{ \log \frac{m(y)}{m_{f,\theta}(y)} \right\}^2 m(y) dy \leq B.$$

If the mixture model is correctly specified, i.e., $m \in \mathbb{M}$, then Assumption 6 is a consequence of Assumption 5 and Jensen’s inequality; see the argument for (20) in Appendix 1. But this assumption seems reasonable even if $m \notin \mathbb{M}$, as it assures that the proposed mixture model (4) is, in a certain sense, close enough to the truth.

Theorem 2. *Under Assumptions 1–6, $K_n(\theta) \rightarrow K^*(\theta)$ almost surely as $n \rightarrow \infty$ for each fixed θ . In addition to Assumptions 1–6, if the infimum in (12) is attained in the interior of \mathbb{F} , and if $\sum_n a_n w_n^2 < \infty$, then $K_n(\theta) - K^*(\theta) = O(a_n^{-1})$ almost surely.*

Proof. See Appendix 1. □

From Theorem 2 we can conclude that $\ell_n(\theta) = -nK^*(\theta) + \ell_{0n} + o(n)$. Therefore for any θ_1, θ_2 with $K^*(\theta_1) \neq K^*(\theta_2)$, the difference $\ell_n(\theta_1) - \ell_n(\theta_2)$ grows linearly in n . Such linear growth in log likelihood differences is the building block of Wald’s famous proof of consistency of maximum likelihood estimators. In our case, we have the following.

Theorem 3. *Suppose Θ is a finite set and Assumptions 1–6 hold. Let $\Theta^* = \{\theta' : K^*(\theta') = \inf_{\Theta} K^*(\theta)\}$. Then $\hat{\theta}_n \in \Theta^*$ almost surely for all sufficiently large n . In particular, if $\Theta^* = \{\theta^*\}$, then $\hat{\theta}_n \rightarrow \theta^*$ almost surely.*

Proof. Let $\Theta^\dagger = \Theta \setminus \Theta^*$. By Theorem 2 and the finiteness of Θ , $\sup_{\Theta^\dagger} \ell_n(\theta) - \ell_n(\theta^*) = -n\{\inf_{\Theta^\dagger} K^*(\theta) - \inf_{\Theta} K^*(\theta)\} + o(n) \rightarrow -\infty$, almost surely. On the other hand, $\ell(\hat{\theta}_n) - \ell_n(\theta^*) \geq 0$ by definition of $\hat{\theta}_n$ and thus $\hat{\theta}_n \notin \Theta^\dagger$ almost surely for all large n . □

While Theorem 3 is limited to finite Θ , our empirical results Section 3.3 suggest that the convergence can be extended to a compact Θ . Toward this, we must show that K_n is either uniformly equi-continuous or convex, a task to be taken up in a future work.

It is unclear whether results similar to Theorems 2 and 3 hold for the profile likelihood L_n^P . Our proofs of these two theorems are based on martingale theory and rely on the fact that the summand $\log m_{i-1,\theta}(Y_i)$ in $L_n^M(\theta)$ is measurable with respect to the σ -algebra generated by Y_1, \dots, Y_i . This does not hold for the summands $\log m_{n,\theta}(Y_i)$ in $\log L_n^P(\theta)$.

3.3 Numerical illustration

Suppose Y_1, \dots, Y_n are modeled as independent observations from the mixture density $m_{f,\sigma}(y) = \int_0^1 p(y \mid \sigma, u) f(u) du$, where $p(y \mid \sigma, u)$ is a $N(u, \sigma^2)$ kernel. The true model, however, is $Y_i = 0.5 + 0.1Z_i$, where Z_1, \dots, Z_n is a random sample from the Student-t distribution with 5 degrees of freedom. The true density $m(y)$ underlying the shifted and scaled Student-t model cannot equal $m_{f,\sigma}(y)$ for any f with support $[0, 1]$ and $\sigma > 0$, because the tails of $m_{f,\sigma}(y)$ decay exponentially while those of $m(y)$ decay polynomially. But despite this model mis-specification, $K_n(\sigma) \rightarrow K^*(\sigma)$ pointwise in σ by Theorem 2.

We approximate $K(m, m_{f,\sigma})$ with its Gaussian quadrature form

$$\int m(y) \log \frac{m(y)}{m_{f,\sigma}(y)} dy \approx \sum_{r=1}^R b_r m(y_r) \log \frac{m(y_r)}{\sum_{j=1}^J a_j p(y_r \mid \sigma, u_j) f(u_j)},$$

which is then numerically optimized over $f(u_1), \dots, f(u_J)$ to obtain an approximation to $K^*(\sigma)$. Here $\{(a_j, u_j) : j = 1, \dots, J\}$ are the Legendre node-weight pairs of order $J = 101$ for the interval $[0, 1]$, and $\{(b_r, y_r) : r = 1, \dots, R\}$ are the same for $[-0.5, 1.5]$ with $R = 101$. Figure 1 shows $K_n(\sigma)$ and the limit $K^*(\sigma)$ for three choices of n , each replicated with 100 independent data sets from the shifted and scaled Student-t distribution. In addition to showing pointwise convergence, Figure 1 also gives an indication of the conjectured convexity of $K_n(\sigma)$.

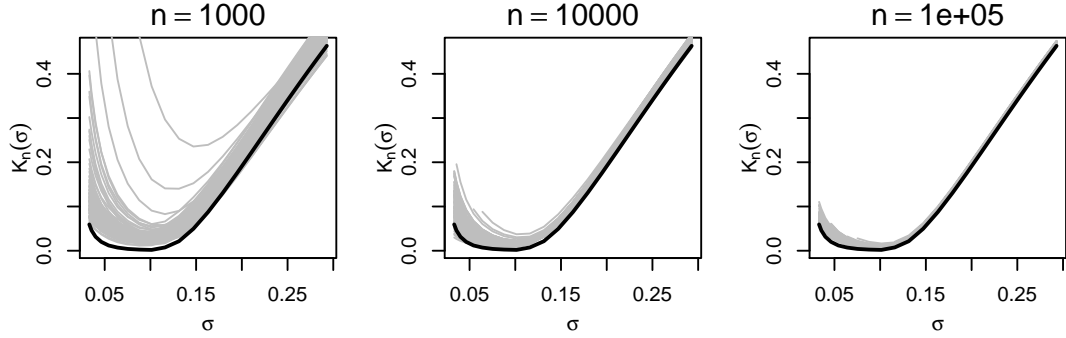


Figure 1: Gray lines show $K_n(\sigma)$ for 100 independent data sets from the mis-specified mixture model in Section 3.3. Black lines show an approximation of the pointwise limit $K^*(\sigma)$.

4 Examples

4.1 Density estimation

Consider density estimation where $p(y | \sigma, u) = N(y | u, \sigma^2)$ is the Gaussian kernel with bandwidth σ playing the role of θ and the mixing density f assumed to have support $[0, 1]$. For the Bayes approach, the distribution function F is modeled as a draw from a Dirichlet process distribution with unit precision parameter and a uniform base measure on $[0, 1]$. For predictive recursion, we take f_0 to be a uniform density on $[0, 1]$, matching the Dirichlet process base measure, and set $w_i = (i + 1)^{-2/3}$. Figure 2 shows $L_n^B(\sigma)$, $L_n^M(\sigma)$ and $L_n^P(\sigma)$ for 12 data sets of size $n = 50$ simulated from the Gaussian mixture model $m_{f,\sigma}(y) = \int p(y | \sigma, u)f(u) du$ for several (σ, f) pairs. The four different mixing distributions were chosen to capture various shapes and characteristics, including discrete, continuous, and a mixture of each. The importance sampling technique of Tokdar et al. (2009) is used to evaluate $L_n^B(\sigma)$; see also MacEachern et al. (1999).

Figure 2 shows the normalized marginal likelihoods for the three different methods. It is clear that $L_n^M(\sigma)$ closely approximates $L_n^B(\sigma)$ in all cases, while $L_n^P(\sigma)$ deviates arbitrarily, sometimes peaking at $\sigma = 0$. The approximation of $L_n^B(\sigma)$ by $L_n^M(\sigma)$ is robust against the smoothness and skewness properties of the underlying mixing distribution f . This is quite striking because the Dirichlet process formulation views f as a discrete distribution while predictive recursion is designed to recover smooth densities. We also note that both the approximate marginal and profile likelihood calculations are orders of magnitude faster than those for Dirichlet process mixture; see Tokdar et al. (2009) for a comparison of run times.

4.2 Random-intercept regression models

In this section we study two regression models along the lines of Tao et al. (1999). In each case, we consider data on a response Y and a vector of predictors $X \in \mathbb{R}^d$ for n subjects each with r replicates. In our first study, Y is a continuous variable and is linked

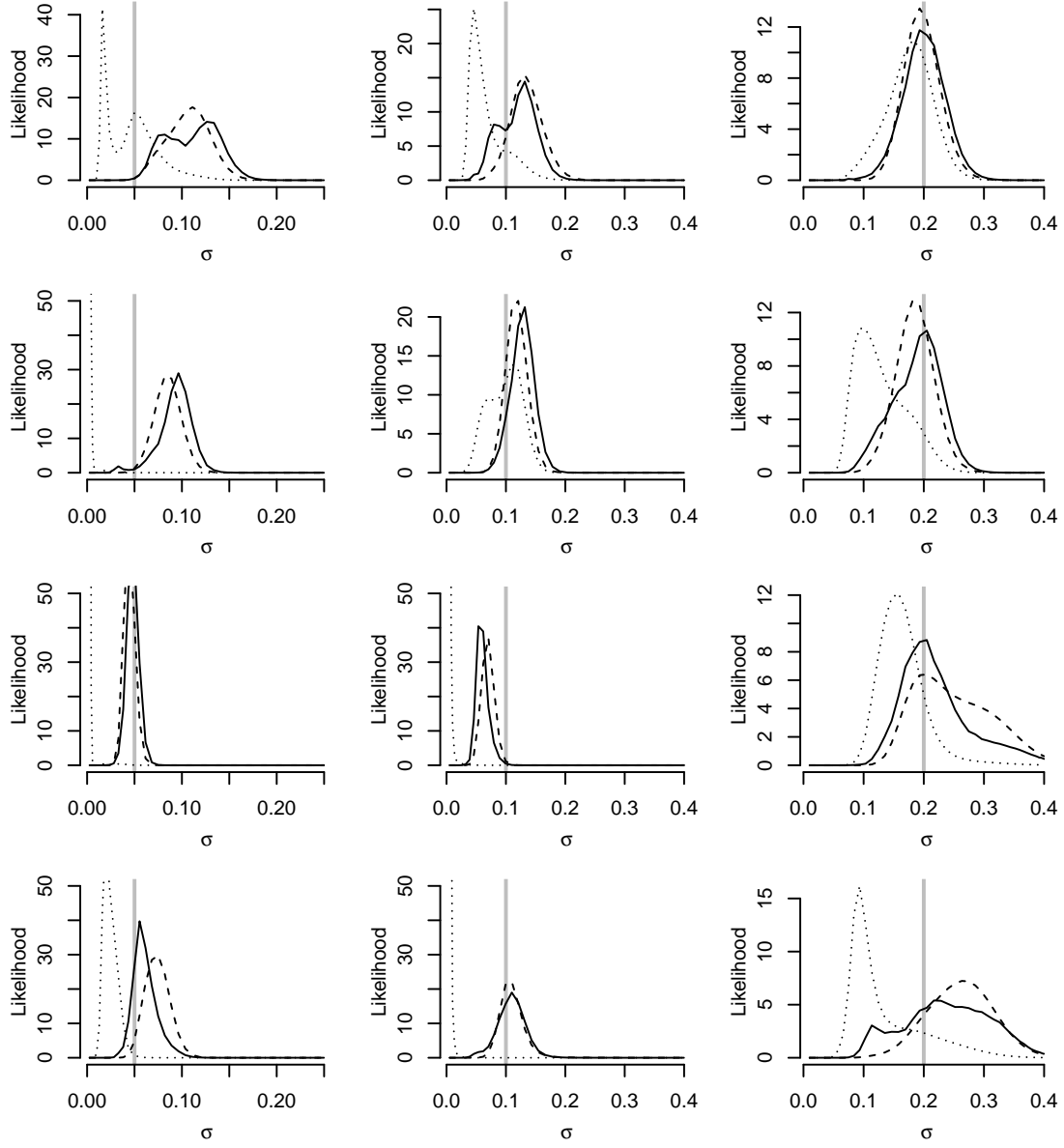


Figure 2: Solid lines show $L_n^B(\sigma)$, dashed lines show $L_n^M(\sigma)$, and dotted lines show $L_n^P(\sigma)$ for the density estimate example in Section 4.1. These curves have been normalized to integrate to 1. The true value of σ is marked by a vertical gray line. The figure's four rows correspond to four different mixing distributions. The first is a Beta(2,6) density; the second is a Beta(10,30) density; the third is an equal mixture of point masses at 1/4 and 3/4; the fourth is an equal mixture of Beta(2,6) and a point mass at 3/4.

to the predictors through the random-intercept linear regression model:

$$Y_{ij} \mid X_{ij} \sim N(U_i + X'_{ij}\beta, \sigma^2), \quad (14)$$

where $i = 1, \dots, n$ indexes subjects, and $j = 1, \dots, r$ indexes replicates. We assume that the Y_{ij} 's are conditionally independent across both i and j . The subject-specific intercepts, the U_i 's, are taken to be independent draws from a probability density f with respect to the Lebesgue measure on an interval $\mathcal{U} = [a, b]$. Write $\theta = (\beta, \sigma^2)$ as the unknown parameter of interest. A related semiparametric Bayes model appears in Bush and MacEachern (1996).

To cast this regression model as the mixture model (4), we assume $X_i = (X_{i1}, \dots, X_{ir})$ are stochastic, sampled independently from a density g over $\mathbb{R}^d \times \dots \times \mathbb{R}^d$. Then (14) says that (X_i, Y_i) , where $Y_i = (Y_{i1}, \dots, Y_{ir})$, are independent observations from a density $m = m_{f,\theta}$ as in (4) with a kernel

$$p(x, y \mid \theta, u) = p_x(y \mid \theta, u)g(x), \quad (15)$$

where the conditional density of Y given x is

$$p_x(y \mid \theta, u) = \frac{1}{(2\pi\sigma^2)^{r/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{j=1}^r (y_j - x'_j\beta - u)^2\right\}.$$

With this setup, θ can be estimated by maximizing the predictive recursion marginal or profile likelihood. The predictor density g need not be estimated: it drops out from the updating equation (2), and so does not affect $f_{i,\theta}$. Consequently, $m_{i,\theta}(x, y) = m_{i,\theta,x}(y)g(x)$, where $m_{i,\theta,x}(y) = \int p_x(y \mid \theta, u) f_{i,\theta}(u) du$ does not involve g , and hence $\log L_n^M(\theta)$ incorporates g only through an additive constant. Also note that by factoring $m(x, y) = m_x(y)g(x)$, one can write $K(m, m_{f,\theta}) = \int K(m_x, m_{f,\theta,x})g(x) dx$ where $m_{f,\theta,x}(y) = \int p_x(y \mid \theta, u) f(u) du$. Thus the (θ^*, f^*) that characterizes the limiting asymptotic properties of $L_n^M(\theta)$ minimizes an average Kullback–Leibler divergence $K(m_x, m_{f,\theta,x})$ of the conditional densities weighted by $g(x)$.

Evaluation of $L_n^M(\theta)$ requires a single pass through the predictive recursion algorithm for each θ , which is computationally inexpensive, even for large n . Optimization, however, requires several evaluations of L_n^M and its gradient. With computational efficiency in mind, we present a version of predictive recursion that also produces ∇L_n^M as a by-product, with no substantial increase in computational cost; see Appendix 2. This gradient algorithm, coupled with any packaged optimization routine, makes for fast semiparametric estimation of θ .

For inference on θ , the exact sampling distribution of $\hat{\theta}$ in (7) will not be available in general, so some sort of approximation is needed. Here we propose a curvature based approximation, where the covariance matrix of $\hat{\theta}$ is estimated by the inverse Hessian $H = \{-\nabla^2 \ell_n(\hat{\theta})\}^{-1}$, readily obtained from the output of the optimization routine. A stipulated $100(1 - \alpha)\%$ confidence interval for θ_j can be obtained by taking $\hat{\theta}_j \pm z_{\alpha/2} h_{jj}$, where z_q denotes the upper q^{th} quantile of the standard normal distribution, and h_{jj} is the j^{th} diagonal element of H .

Table 1 reports the performance of the predictive recursion-based marginal and profile likelihood estimates averaged over 500 datasets generated from model (14) with $n = 50$

or 500, $r = 4$, $d = 2$, $\beta = (2, 5)'$, and $\sigma = 2$. The covariates X_{ij1} 's are independent $N(0, 1)$ and $X_{ij2} = J_i + 0.1Z_{ij}$, with J_i 's satisfying $P(J_i = 0) = P(J_i = 1) = 0.5$ and Z_{ij} 's independent $N(0, 1)$. In this case, X_{ij1} is a within-subject covariate while X_{ij2} , which is roughly constant in j , acts like a subject-specific covariate. Three choices of f are considered: $N(0, 2^2)$, a shifted exponential distribution with rate 0.5 and support $(-2, \infty)$, and a discrete uniform supported on ± 2 . Each choice of f has mean zero and variance 4. See Tao et al. (1999) for more details on these choices. For comparison to a parametric model fit, we also include a likelihood-based method that assumes the mixing distribution is Gaussian with unknown parameters.

Performance of each estimate is measured by the root mean-square error in estimating each component of θ . We also include average coverage of a stipulated 95% confidence interval for each estimating method constructed as described above. Predictive recursion is run with $w_i = (1 + i)^{-2/3}$, $\mathcal{U} = \bar{Y} \pm 3S_Y$ where \bar{Y} and S_Y are the mean and standard deviation of the full data $\{Y_{ij}\}$, and f_0 the uniform density over \mathcal{U} . In this example, the predictive recursion-based marginal and profile likelihood methods perform similarly. Both are competitive with the parametric Gaussian method when f is indeed Gaussian, and are better for non-Gaussian f . The relative similarity between the marginal and the profile likelihood methods, unlike what we observed in the density estimation example before, can be explained by noting that for model (14), the data is informative about the parameter σ due to availability of replicates. We note that although the coverage probabilities are generally close to the stipulated 95% level, there are some noticeable differences. First, the marginal and profile likelihood coverage probabilities for β_2 are off the mark in the small n case. That X_{ij2} is partially confounded with the group structure is one potential explanation for this phenomenon. Second, in estimating σ for the discrete uniform model, which lies in the boundary of \mathbb{F} , the coverage falls dangerously low, even for large n . For such boundary cases, bootstrap confidence intervals might be more appropriate; see Section 5.

In our second example, we retain much of the above setting but consider a binary Y for which model (14) is adapted to the a random-intercept logistic regression model:

$$\text{logit}\{P(Y_{ij} = 1)\} = U_i + X'_{ij}\beta, \quad (16)$$

where U_i 's are independent draws from a uniform density f on $\mathcal{U} = [-8, 8]$ and $\theta = (\beta_1, \beta_2)$ is unknown. This model corresponds to (4) through (15) with an appropriate choice of $p_x(y|\theta, u)$. The last four columns of Table 1 reports the performance of maximum likelihood estimates of θ based on the predictive recursion marginal and profile likelihood and the parametric Gaussian set, with same choices for β_1 , β_2 , n , r , f , w_i , and $\{X_{ij}\}$ as in our first study. For $n = 500$, all methods perform similarly; for $n = 50$ the proposed marginal likelihood approach is better in terms of both estimation accuracy and coverage.

The marginal and profile likelihood methods used above inherit the order-dependence characteristic of predictive recursion. That is, $L_n^M(\theta)$ and $L_n^P(\theta)$ depend on the order in which the data enter the recursion (2). The order is not important asymptotically, but finite sample behaviors may be sensitive to the specific order used. A simple remedy to this is to apply predictive recursion on M random permutations of the data and average the M resulting marginal and profile likelihood functions prior to optimization. Permutation-averaging in the two regression examples gave only marginally better results compared to what is shown in Table 1.

f	Method	Study I						Study II			
		RMSE			Coverage			RMSE		Coverage	
		β_1	β_2	σ	β_1	β_2	σ	β_1	β_2	β_1	β_2
		$n = 50$						$n = 50$			
Gaussian	Gaussian	0.16	0.60	0.12	95	95	94	1.02	2.48	88	86
	Marginal	0.16	0.68	0.12	95	86	94	0.61	1.58	95	91
	Profile	0.16	0.71	0.12	94	89	92	0.65	1.78	95	90
Exponential	Gaussian	0.17	0.61	0.11	93	94	95	0.86	2.13	90	91
	Marginal	0.17	0.52	0.11	94	91	95	0.64	1.66	96	96
	Profile	0.17	0.50	0.12	94	92	94	0.67	1.76	96	96
Uniform	Gaussian	0.15	0.58	0.11	96	95	95	0.94	2.42	87	88
	Marginal	0.14	0.36	0.11	96	97	93	0.82	1.92	92	92
	Profile	0.14	0.34	0.12	96	97	90	0.89	2.14	92	93
		$n = 500$						$n = 500$			
Gaussian	Gaussian	0.05	0.18	0.04	96	95	95	0.17	0.43	88	85
	Marginal	0.05	0.19	0.04	96	94	95	0.15	0.39	93	93
	Profile	0.05	0.19	0.04	96	94	95	0.16	0.41	93	95
Exponential	Gaussian	0.05	0.18	0.04	93	96	95	0.15	0.41	90	87
	Marginal	0.05	0.15	0.04	95	94	95	0.15	0.32	96	96
	Profile	0.05	0.14	0.04	95	95	94	0.15	0.34	95	95
Uniform	Gaussian	0.05	0.19	0.04	96	94	94	0.18	0.50	88	84
	Marginal	0.05	0.11	0.05	94	95	80	0.20	0.52	91	89
	Profile	0.05	0.11	0.05	95	94	84	0.21	0.54	90	86

Table 1: Comparison of the predictive recursion-based marginal and profile likelihood methods along with a parametric Gaussian model for parameter estimation in the random-intercept regression models in Section 4.2. RMSE is root mean-square error. Coverage probabilities are multiplied by 100. Studies I and II denote the linear and logistic models, respectively.

4.3 Multiple testing in time series

Let $Y = \{Y(t) : t = 1, 2, \dots, T\}$ denote a discrete-time stochastic process under a first-order autoregressive model; i.e., $Y(0)$ is normal with mean 0 and variance $\sigma^2(1 - \phi)^{-1}$, and

$$Y(t) = \xi + \phi Y(t-1) + \sigma \varepsilon(t), \quad \varepsilon(t) \sim N(0, 1), \quad t = 1, 2, \dots, T,$$

for $\xi \in \mathbb{R}$, $\phi \in (-1, 1)$ and $\sigma^2 > 0$. In other words, Y has a T -dimensional normal distribution with mean $\xi 1_T = (\xi, \dots, \xi)' \in \mathbb{R}^T$ and $T \times T$ covariance matrix Σ_u of the form

$$\Sigma_{u,jk} = \frac{\sigma^2}{1 - \phi} \phi^{|j-k|}, \quad u = (\sigma^2, \phi) \in \mathcal{U} \subset (0, \infty) \times (-1, 1).$$

For a sample Y_1, \dots, Y_n of similar processes, consider a mixture of simple first-order autoregressive models, namely,

$$Y_i \mid (\xi_i, U_i) \sim N(\xi_i 1_T, \Sigma_{U_i}), \quad U_i \sim f(u), \quad \xi_i \mid \theta \sim \theta \langle 0 \rangle + (1 - \theta) N(0, 1), \quad (17)$$

where independence is assumed throughout, and $\langle 0 \rangle$ denotes a degenerate distribution at 0. It shall be further assumed that θ is close to 1, so that Y_1, \dots, Y_n are sparse in the sense that most have mean zero. The goal is to identify those processes with non-zero mean trajectories.

Model (17) is similar to that which appears in Scott (2009). He considers a financial time-series application in which $Y_i(t)$ is related to the i^{th} firm's return on assets for year t . Scott (2009) takes the $Y_i(t)$ to be the standardized residuals obtained when return on assets is regressed on several important factors. Therefore, firms whose Y -process has zero mean are ordinary, those with non-zero mean are somehow extraordinary. The goal is to flag those extraordinary firms. Scott (2009) considers a nonparametric Bayes model where f is sampled from a Dirichlet process distribution, and the signal trajectories are constant zero with probability θ or sampled from a Gaussian process with probability $1 - \theta$. Model (17) is a special case in which the signal trajectories are restricted to constant functions. But the assumption of constant paths is not critical. Indeed, the marginal likelihood procedure described below can be extended to handle signal trajectories with a relatively low-dimensional parametrization.

From model (17), it is clear that Y_1, \dots, Y_n are independent with common mixture density

$$\begin{aligned} m_{f,\theta}(y) &= \int_{\mathcal{U}} \int_{-\infty}^{\infty} N(y \mid \xi 1, \Sigma_u) \{ \theta \langle 0 \rangle(\xi) + (1 - \theta) N(\xi \mid 0, 1) \} d\xi f(u) du \\ &= \int_{\mathcal{U}} \{ \theta N(y \mid 0, \Sigma_u) + (1 - \theta) N(y \mid 0, \Sigma_u + 1_T 1_T') \} f(u) du. \end{aligned} \quad (18)$$

Here, with the particular choice of parametric mixing distribution, ξ can be integrated out analytically, leaving only a mixture over u . Predictive recursion marginal likelihood is used to estimate θ and f and, in turn, these estimates are used to classify the sample paths Y_1, \dots, Y_n .

If θ and f were known, the Bayes oracle rule for classifying Y_i as null or non-null is as follows: for 0–1 loss, conclude Y_i is non-null if

$$\theta_i = \frac{\theta \int N(Y_i \mid 0, \Sigma_u) f(u) du}{m_{f,\theta}(Y_i)} \leq 0.5.$$

	Method	Mean	SD	FDR	MP
Dense, $\theta = 0.75$	Plug-in	0.750	0.008	0.085	0.109
	Oracle	—	—	0.083	0.108
Sparse, $\theta = 0.95$	Plug-in	0.949	0.004	0.082	0.026
	Oracle	—	—	0.073	0.026

Table 2: Summary of inference in the autoregressive process mixture model simulations for the plug-in empirical Bayes and Bayes oracle methods. Mean denotes the simulation mean of $\hat{\theta}$ and SD is its standard deviation; FDR is the observed false discovery rate of the test and MP is its misclassification probability.

Since θ and f are unknown, we mimic the Bayes oracle classifier with the plug-in estimate

$$\hat{\theta}_i = \frac{\hat{\theta} \int N(Y_i | 0, \Sigma_u) f_{n, \hat{\theta}}(u) du}{m_{n, \hat{\theta}}(Y_i)}, \quad (19)$$

which is an estimate of the local false discovery rate (Efron 2004, 2008; Sun and Cai 2007), and can be viewed as an empirical Bayes approximation of the posterior inclusion probabilities in Scott (2009), there obtained by Markov chain Monte Carlo.

For illustration, we simulate 100 datasets from the model (17), with $n = 5000$ and $T = 50$, and compare the performance of our empirical Bayes classifier (19) to the Bayes oracle. We consider both a dense case, $\theta = 0.75$, and a sparse case, $\theta = 0.95$. In both cases the true mixing distribution $f(u) = g_1(\phi)g_2(\sigma^2)$ is a product of shifted and scaled beta densities. Predictive recursion marginal likelihood is applied, averaging over 25 random permutations, to estimate θ and a summary of the estimates for both the dense and sparse cases can be found in Table 2. We find that the estimates of θ are unbiased with relatively small standard errors. Histograms, not displayed, show roughly symmetric distributions centered at the true θ .

For the testing/classification problem, we consider two measures of performance: false discovery rate and misclassification probability. Empirical estimates of these two quantities appear in Table 2 for both the plug-in and Bayes oracle rules, and both dense and sparse cases. The difference between false discovery rates is negligible in both the dense and sparse cases. Furthermore, the misclassification probability for the predictive recursion-based empirical Bayes classifier is just slightly higher than that of the Bayes oracle, suggesting that the latter mimics the Bayes oracle classifier very well.

5 Discussion

The focus in this paper is on frequentist semiparametric inference in mixture models with a predictive recursion-based approximation to a Dirichlet process mixture marginal likelihood. But a Bayesian version can be developed without much additional effort. In particular, information gathered from maximizing $L_n^M(\theta)$ in (6) can be used to construct efficient importance samplers to carry out posterior computation, i.e., the maximizer and Hessian matrix of $L_n^M(\theta)$ can be used to construct Gaussian or heavy-tailed Student-t importance densities for θ (Geweke 1989).

The choice of weights $\{w_n : n \geq 1\}$ for predictive recursion remains an important open problem. Convergence theory gives only minimal guidelines but, in our experience, the finite sample performance is relatively robust to the choice of weights. In this paper, we have employed the theoretically ideal weights $w_i = (i + 1)^{-\gamma}$, with $\gamma = 2/3$, based on the rate in Theorem 1. An alternative is to let the exponent γ be an additional tuning parameter to maximize the approximate marginal likelihood L_n^M over (Tao et al. 1999). It is not yet clear, however, if the convergence theorems of Martin and Tokdar (2009) can cover data-dependent weight sequences. In the random-intercept regression problems of Section 4.2, the approach with estimated γ gave similar results, not reported, to the that with fixed $\gamma = 2/3$.

The use of Hessian-based approximations of the sampling distribution of the predictive recursion marginal likelihood estimate $\hat{\theta}$ is based on a theory of asymptotic normality which is not yet available. An alternative to the Hessian-based approach is a basic bootstrap. Empirical results, not presented here, indicate that bootstrap-based confidence intervals have good coverage properties, but progress on the validity of the bootstrap for semiparametric problems has only just recently been made (Cheng and Huang 2010). Since $\log L_n^M(\theta)$ is not an empirical processes, these first results do not directly apply in our context. It is our experience that both of these approaches for approximating the sampling distribution of $\hat{\theta}_n$ are successful, but we leave theoretical verification of their validity to future research.

Acknowledgments

The authors are grateful to Professor J. K. Ghosh for many helpful discussions, and to the Associate Editor and two referees for a number of invaluable suggestions.

Appendix 1: Proof of Theorem 2

Fix θ and define the sequence of random variables $Z_i = Z_i(\theta)$ as

$$Z_i = \log \frac{m(Y_i)}{m_{i-1,\theta}(Y_i)} - K(m, m_{i-1,\theta}), \quad i \geq 1$$

and note that $E(Z_i | \mathcal{A}_{i-1}) = 0$, where $\mathcal{A}_{i-1} = \sigma(Y_i, \dots, Y_{i-1})$. Therefore, $\{(Z_i, \mathcal{A}_i) : i \geq 1\}$ forms a zero mean martingale sequence. Next, let $m_{f,\theta}$ be the mixture density closest to m in the Kullback–Leibler sense. Then we can write

$$\begin{aligned} E(Z_i^2 | \mathcal{A}_{i-1}) &\leq \int \left\{ \log \frac{m(y)}{m_{i-1,\theta}(y)} \right\}^2 m(y) dy \\ &= \int \left\{ \log \frac{m_{f,\theta}(y)}{m_{i-1,\theta}(y)} + \log \frac{m(y)}{m_{f,\theta}(y)} \right\}^2 m(y) dy \\ &\leq 2 \int \left\{ \log \frac{m_{f,\theta}(y)}{m_{i-1,\theta}(y)} \right\}^2 m(y) dy + 2 \int \left\{ \log \frac{m(y)}{m_{f,\theta}(y)} \right\}^2 m(y) dy \\ &= 2T_1 + 2T_2. \end{aligned}$$

The second term, T_2 , is bounded by a constant B according to Assumption 6. For the first term, let $\mathcal{Y}_0 = \{y : m_{f,\theta}(y) < m_{i-1,\theta}(y)\}$. By properties of the logarithm we get

$$\begin{aligned}
T_1 &= \int \left\{ \log \frac{m_{f,\theta}(y)}{m_{i-1,\theta}(y)} \right\}^2 m(y) dy \\
&= \int_{\mathcal{Y}_0} \left\{ \log \frac{m_{i-1,\theta}(y)}{m_{f,\theta}(y)} \right\}^2 m(y) dy + \int_{\mathcal{Y}_0^c} \left\{ \log \frac{m_{f,\theta}(y)}{m_{i-1,\theta}(y)} \right\}^2 m(y) dy \\
&\leq \int_{\mathcal{Y}_0} \left\{ \frac{m_{i-1,\theta}(y)}{m_{f,\theta}(y)} - 1 \right\}^2 m(y) dy + \int_{\mathcal{Y}_0^c} \left\{ \frac{m_{f,\theta}(y)}{m_{i-1,\theta}(y)} - 1 \right\}^2 m(y) dy \\
&\leq 2 + \int \left[\left\{ \frac{m_{i-1,\theta}(y)}{m_{f,\theta}(y)} \right\}^2 + \left\{ \frac{m_{f,\theta}(y)}{m_{i-1,\theta}(y)} \right\}^2 \right] m(y) dy \\
&\leq 2 + 2 \sup_{u_1, u_2} \int \left\{ \frac{p(y|u_1, \theta)}{p(y|u_2, \theta)} \right\}^2 m(y) dy,
\end{aligned} \tag{20}$$

where (20) follows by two applications of Jensen's inequality, one to the mapping $x \mapsto 1/x$ and one to the mapping $x \mapsto x^2$. The last term is bounded according to Assumption 5. Therefore, $\mathbb{E}(Z_i^2 \mid \mathcal{A}_{i-1})$ is uniformly bounded by a constant M and, consequently, $v_n = \sum_{i=1}^n \mathbb{E}(Z_i^2 \mid \mathcal{A}_{i-1}) \leq Mn$.

Set $b_n = n/c_n$, where $c_n = a_n$ or $c_n \equiv 1$, depending on whether or not the conditions of the second part of the theorem hold. It is clear that b_n grows faster than $n^{1/2}$, but no faster than n , which implies

$$\frac{v_n^{1/2}}{b_n(\log \log b_n)^{-1/2}} \leq \frac{(Mn \log \log b_n)^{1/2}}{b_n} \rightarrow 0. \tag{21}$$

Furthermore, by Markov's inequality, we have, with probability 1,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|Z_n| > \frac{b_n}{\log \log b_n} \mid \mathcal{A}_{n-1}\right) \leq M \sum_{n=1}^{\infty} \frac{(\log \log b_n)^2}{b_n^2} < \infty. \tag{22}$$

In light of (21) and (22), it now follows from Corollary 2 of Teicher (1998) that $b_n^{-1} \sum_{i=1}^n Z_i \rightarrow 0$ almost surely. Therefore, we can conclude that, with probability 1,

$$\begin{aligned}
&c_n \left| K_n(\theta) - \frac{1}{n} \sum_{i=1}^n K(m, m_{i-1,\theta}) \right| \\
&= \left| c_n \{K_n(\theta) - K^*(\theta)\} - \frac{c_n}{n} \sum_{i=1}^n \{K(m, m_{i-1,\theta}) - K^*(\theta)\} \right| \rightarrow 0.
\end{aligned} \tag{23}$$

If $c_n \equiv 1$, then the result follows from Theorem 1 and Cesaro's theorem. If $c_n = a_n$, write $\kappa_i = K(m, m_{i-1,\theta}) - K^*(\theta)$ and note that $\kappa_i \geq 0$. Summation-by-parts and monotonicity of the weights w_i yield the following inequality:

$$a_n \sum_{i=1}^n \kappa_i = \sum_{i=1}^n a_i \kappa_i + \sum_{i=1}^{n-1} w_{i+1} \left(\sum_{j=1}^i \kappa_j \right) \leq \sum_{i=1}^n a_i \kappa_i + \sum_{i=1}^n \left(\sum_{j=1}^i w_j \kappa_j \right). \tag{24}$$

Let $S_i = \sum_{j=1}^i w_j \kappa_j$ for $i \geq 1$. Lemma 4.4 of Martin and Tokdar (2009) shows that the limit S_∞ is finite almost surely. Dividing through (24) by n and applying Cesaro's

theorem shows that the right-hand side is positive and bounded by S_∞ almost surely for large n . This observation, together with (23), implies $a_n\{K_n(\theta) - K^*(\theta)\}$ is almost surely bounded, proving the theorem.

Appendix 2: Predictive recursion gradient algorithm

Here we present a version of the predictive recursion that gives the gradient of $\ell_n(\theta) = \log L_n^M(\theta)$ as a by-product. Let $\lambda_i(\theta) = m_{i-1,\theta}(Y_i)$; then $\nabla \ell_n(\theta) = \sum_{i=1}^n \nabla \log \lambda_i(\theta)$. For a function $g(\theta, u)$, the notation $\nabla g(\theta, u)$ means the gradient of g with respect to θ , pointwise in u .

As in the original algorithm, the user must be able to evaluate the kernel $p(Y_i | \theta, u)$ at each Y_i for any pair (θ, u) . Furthermore, for this modification, the user must also be able to evaluate $\nabla p(Y_i | \theta, u)$.

1. Start with the user-defined $f_{0,\theta}(u)$ and compute $\nabla f_{0,\theta}(u)$.

2. For $i = 1, \dots, n$, repeat the following three steps:

(a) Set $g(\theta, u) = p(Y_i | \theta, u)$, $\nabla g(\theta, u) = \nabla p(Y_i | \theta, u)$, and

$$G(\theta, u) = g(\theta, u)\nabla f_{i-1,\theta}(u) + \nabla g(\theta, u)f_{i-1,\theta}(u).$$

(b) Compute $\lambda_i(\theta) = \int g(\theta, u)f_{i-1,\theta}(u) d\mu(u)$ and $\nabla \log \lambda_i(\theta) = \int G(\theta, u) d\mu(u)/\lambda_i(\theta)$.

(c) Update

$$f_{i,\theta}(u) = (1 - w_i)f_{i-1,\theta}(u) + w_i \frac{g(\theta, u)f_{i-1,\theta}(u)}{\lambda_i(\theta)},$$

$$\nabla f_{i,\theta}(u) = (1 - w_i)\nabla f_{i-1,\theta}(u) + w_i \left\{ \frac{G(\theta, u) - g(\theta, u)f_{i-1,\theta}(u)\nabla \log \lambda_i(\theta)}{\lambda_i(\theta)} \right\}.$$

3. Return $f_{n,\theta}(u)$, $m_{n,\theta}(y)$, and $\sum_{i=1}^n \nabla \log \lambda_i(\theta)$.

References

- Antoniak, C. (1974), “Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems,” *Ann. Statist.*, 2, 1152–1174.
- Blackwell, D. and MacQueen, J. B. (1973), “Ferguson distributions via Pólya urn schemes,” *Ann. Statist.*, 1, 353–355.
- Bush, C. A. and MacEachern, S. N. (1996), “A semiparametric Bayesian model for randomised block designs,” *Biometrika*, 83, 275–285.
- Cheng, G. and Huang, J. (2010), “Bootstrap consistency for general semiparametric M-estimation,” *Ann. Statist.*, 38, 2884–2915.
- Efron, B. (2004), “Large-scale simultaneous hypothesis testing: the choice of a null hypothesis,” *J. Amer. Statist. Assoc.*, 99, 96–104.

- (2008), “Microarrays, empirical Bayes and the two-groups model,” *Statist. Sci.*, 23, 1–22.
- Ferguson, T. S. (1973), “A Bayesian analysis of some nonparametric problems,” *Ann. Statist.*, 1, 209–230.
- Geweke, J. (1989), “Bayesian inference in econometric models using Monte Carlo integration,” *Econometrica*, 57, 1317–1339.
- Ghosh, J. K. and Ramamoorthi, R. V. (2003), *Bayesian Nonparametrics*, New York: Springer-Verlag.
- Ghosh, J. K. and Tokdar, S. T. (2006), “Convergence and consistency of Newton’s algorithm for estimating mixing distribution,” in *Frontiers in statistics*, eds. Fan, J. and Kou, H., London: Imp. Coll. Press, pp. 429–443.
- Jefferys, W. and Berger, J. (1992), “Ockham’s razor and Bayesian analysis,” *American Scientist*, 80, 64–72.
- MacEachern, S. N., Clyde, M., and Liu, J. S. (1999), “Sequential importance sampling for nonparametric Bayes models: the next generation,” *Canad. J. Statist.*, 27, 251–267.
- Martin, R. and Ghosh, J. K. (2008), “Stochastic approximation and Newton’s estimate of a mixing distribution,” *Statist. Sci.*, 23, 365–382.
- Martin, R. and Tokdar, S. T. (2009), “Asymptotic properties of predictive recursion: robustness and rate of convergence,” *Electron. J. Stat.*, 3, 1455–1472.
- Newton, M. A. (2002), “On a nonparametric recursive estimator of the mixing distribution,” *Sankhyā Ser. A*, 64, 306–322.
- Newton, M. A., Quintana, F. A., and Zhang, Y. (1998), “Nonparametric Bayes methods using predictive updating,” in *Practical nonparametric and semiparametric Bayesian statistics*, eds. Dey, D., Müller, P., and Sinha, D., New York: Springer, vol. 133 of *Lecture Notes in Statist.*, pp. 45–61.
- Newton, M. A. and Zhang, Y. (1999), “A recursive algorithm for nonparametric analysis with missing data,” *Biometrika*, 86, 15–26.
- Quintana, F. A. and Newton, M. A. (2000), “Computational aspects of nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences,” *J. Comput. Graph. Statist.*, 9, 711–737.
- Scott, J. G. (2009), “Nonparametric Bayesian multiple testing for longitudinal performance stratification,” *Ann. Appl. Statist.*, 3, 1655–1674.
- Sun, W. and Cai, T. T. (2007), “Oracle and adaptive compound decision rules for false discovery rate control,” *J. Amer. Statist. Assoc.*, 102, 901–912.
- Tao, H., Palta, M., Yandell, B. S., and Newton, M. A. (1999), “An estimation method for the semiparametric mixed effects model,” *Biometrics*, 55, 102–110.

- Teicher, H. (1998), “Strong laws for martingale differences and independent random variables,” *J. Theoret. Probab.*, 11, 979–995.
- Tierney, L. and Kadane, J. B. (1986), “Accurate approximations for posterior moments and marginal densities,” *J. Amer. Statist. Assoc.*, 81, 82–86.
- Tokdar, S. T., Martin, R., and Ghosh, J. K. (2009), “Consistency of a recursive estimate of mixing distributions,” *Ann. Statist.*, 37, 2502–2522.
- Wald, A. (1949), “Note on the consistency of the maximum likelihood estimate,” *Ann. Math. Statist.*, 20, 595–601.